文学研究論集
第47号　2017. 9

計量文体論の手法を用いた文芸テキスト分類
――James Tiptree, Jr. と Ernest Hemingway――

Quantitative Authorship Attribution of Two Writers
――A Case Study for James Tiptree, Jr. and Ernest Hemingway――

博士後期課程　英文学専攻　2016年度入学
木　村　美　紀
KIMURA Miki

【論文要旨】

本研究では，Silverberg（1975）や小谷（1994）などの文芸批評においてその文体の類似性が指摘されている James Tiptree, Jr. 作品群と Ernest Hemingway 作品群の文体を計量文体論の手法を用いて検証する。本研究では，Alice Bradley Sheldon の男性名義である James Tiptree, Jr. 作品群67作品を用いて Ernest Hemingway 作品群の文体比較を行うと同時に，分析に使用した指標の有効性・分類手法の有効性を検証していく。分析に用いた統計手法は Baayen（2008）で分類感度の高さが示されているサポートベクターマシン（SVM）と金・村上（2007）において，既存の分類法に比べて分類感度が高いと結論付けられているランダムフォレスト（RF）を採用した。指標としては，Hirst and Feiguina（2007）や Hou and Jiang（2016）で使用されその有効性が示されている品詞分布を採用した。試行の結果，分類正確率は基準を有意に上回っており Ernest Hemingway 作品群と James Tiptree, Jr. 作品群の文体の類似性はとらえることができなかった。James Tiptree, Jr. 名義作品群と Ernest Hemingway 作品群を分類する際に有効である指標は Hirst and Feiguina（2007）で用いられている品詞の bigram であり，分類に有効であった統計手法は SVM であるということが判明した。

【キーワード】 James Tiptree, Jr.，ランダムフォレスト，サポートベクターマシン，著者推定，品詞分布

研究論集委員会　　受付日　2017年 4 月21日　　承認日　2017年 5 月22日

1. はじめに

本研究では，計量文体論の手法を用いて，文芸批評上比較されることの多い James Tiptree, Jr. 作品群と Ernest Hemingway 作品群のテキスト分類を行う。Alice Bradley Sheldon（1915–1987）は James Tiptree, Jr. と Raccoona Sheldon という男女 2 名義を使用しながら正体不明・性別不明の作家として約20年間著作活動を行っていた作家である。主に短編の SF 小説を執筆していた作家であり，文芸批評において Alice Bradley Sheldon の男性名義である James Tiptree, Jr. 名義作品群は Ernest Hemingway 作品群と比較されることが多い。

本研究では，論文執筆者自身が構築した Alice Bradley Sheldon 全72作品（延べ865,802語）を収録したコーパスから James Tiptree, Jr. 作品群のみを抽出し，Ernest Hemingway 作品群との文体比較を行う。Alice Bradley Sheldon という作家は1967年に52歳で James Tiptree, Jr. として作家デビューして以来約10年間にわたり正体不明・性別不明の作家として執筆活動を行っていたため，この作家に関する正体・性別に関する憶測に基づいた文芸批評が数多く存在している。特に Silverberg（1975）や小谷（1994），Larbalestier（2002）に代表されるように，Ernest Hemingway に言及しながら James Tiptree, Jr. 作品群の文体の男性性を論じている文芸批評が多い。本稿では，このように文芸批評で比較されることの多い James Tiptree, Jr. 作品群と Ernest Hemingway 作品群の文体の相違を計量文体論の手法を用いて視覚化する。

分析手法としては，Baayen（2008）において，他の分類手法に比べて分類感度が高いことが多いとされるサポートベクターマシン（SVM）と，金・村上（2007）において様々なジャンルのテキスト分類において，既存の分類法に比べて分類感度が高いと結論付けられているランダムフォレスト（RF）を用いた。分類感度が高いと結論付けられているこれら 2 種類の手法を用いて，James Tiptree, Jr. 作品群と Ernest Hemingway 作品群の計量文体分析を行ったあと，この本研究に用いたデータセットに対してどちらの統計手法が分類に有効であるのか評価を行っていく。分析に用いた指標としては，Hou and Jiang（2016）や Hirst and Feiguina（2007）においてその有効性が示され，木村（2017a）で Alice Bradley Sheldon 全72作品と Ernest Hemingway 作品群69作品の分類に際しその有効性が示されている品詞分布を採用した。

2. 先行研究

2.1 文芸批評における評価

Alice Bradley Sheldon の James Tiptree, Jr. 名義作品群に関する有名な文芸批評として Silverberg（1975）が挙げられる。Silverberg（1975 : vii）では，“It has been suggested that Tiptree is female, a theory that I find absurd, for there is to me something ineluctably masculine about Tiptree’s writing. I don’t think the novels of Jane Austen could have been written by a man nor the stories of Ernest Hemingway by a woman, and in the same way I believe the author of the James Tip-

tree stories is male.'' と，「James Tiptree, Jr. 作品群の著者は男性である」と結論付けている。また，Silverberg（1975：xiv）では，''So, then, James Tiptree—a man of 50 or 55, I guess, possibly unmarried, fond of outdoor life, restless in his everyday existence, a man who has seen much of the world and understands it.'' のように James Tiptree という正体不明・性別不明作家の人物像について「50歳から55歳の男性，未婚」といったような詳細なプロファイリングを行っている。

また，Lefanu（1989：126）においては，'''The masculine manner' of Tiptree's style is cunning contrivance that reveals, first, the limitations of a machismo-oriented culture and the limitations of science fiction when that oriented culture is incorporated unquestioningly into its fictive conventions.'' というように，Silverberg（1975）同様，James Tiptree, Jr. 名義作品群に関する男性性に言及している。さらに，Larbalestier（2002：182）では ''Here are the bare bones of the fable of the woman, Alice Sheldon, writing sf as a man, James Tiptree Jr., who 'proved' there was no essential difference between the way men and women write, but only in the way that what they write is actually read.'' というように男性的な文体・女性的な文体ということに関して評価を与えている。また，Larbalestier（2002：182）では，Robert Silverberg（1975）に言及しながら，''Tiptree's Hemingwayesque ineluctable masculinity'' というように，James Tiptree, Jr. 作品群と Ernest Hemingway 作品群の関係性を指摘している。

日本における Alice Bradley Sheldon 研究の第一人者である小谷（1994: 53）では，「ジェイムズ・ティプトリー・ジュニアなる作家は，（中略）その華麗な文体，ヘミングウェイを思わせるマッチョな作風で一躍 SF 界を魅了した。時代はフェミニズム SF 華やかなりし頃，そのなかでこの著者不明＝正体不明の作品は，その作風から，手堅い稀有の才能を持つ男性新人作家の書いたものと判断されていた。」といったように，Ernest Hemingway に言及しながら，Alice Bradley Sheldon 作品に関して，特に James Tiptree, Jr. 名義の作品の作風・文体に関して主観的な評価を行っている。

2.2 計量文体論の手法を用いた研究

2.2.1 Kimura（2016）

Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群に対し計量文体論手法を用いて分類を行った先行研究としては，論文執筆者自身が実施した Kimura（2016）が挙げられる。Kimura（2016）では，その実績と有効性から Burrows（1987），Burrows（1992），Burrows and Hassal（1988），Rybicky（2015）や，Hou and Jiang（2016）でなどで採用されている「高頻度語彙上位50語」と「品詞分布」を指標として採用した。統計手法としては，教師なしの分類手法であるマルチスケールブートストラップ法に基づくクラスター分析と主成分分析（PCA），教師ありの分類手法である判別分析と SVM を用いた。これらの統計手法を用いて，Alice Bradley Sheldon 男女 2 名義（James Tiptree, Jr. と Raccoona Sheldon）作品群と Ernest Hemingway 作品群の 3 カテゴリ

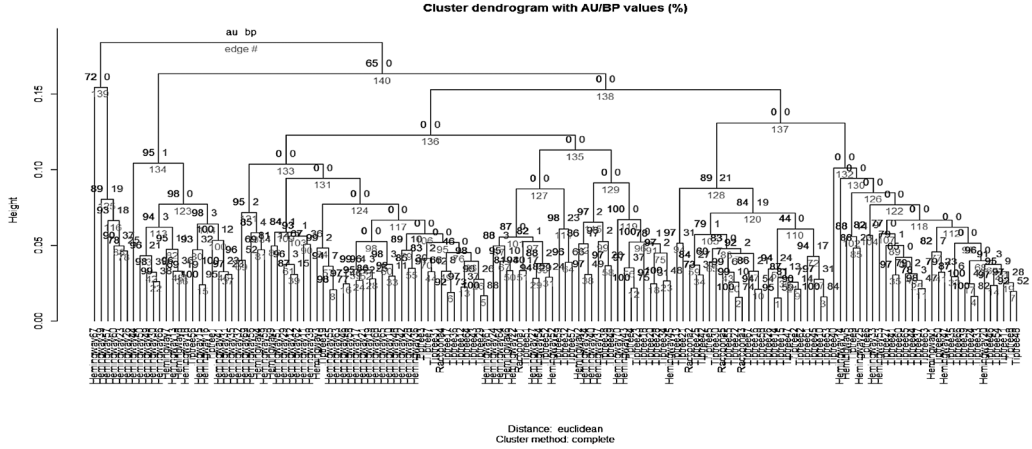

図1　品詞分布を指標として採用したクラスター分析デンドログラム

ーでの分類を検証した。

高頻度語彙上位50語を用いて検証を行ったところ，前述の4種類での統計手法ではAlice Bradley Sheldon作品群における男女2名義での文体差は確認できなかった。一方Alice Bradley Sheldon作品群とErnest Hemingway作品群は完全に分離が可能であった。教師ありの分類手法である判別分析SVMでは，Alice Bradley Sheldonにおける名義での分類は成功せず，Alice Bradley Sheldon作品群とErnest Hemingway作品群の分類は成功した。

また，品詞分布を用いて検証を行ったところ，教師なしの分類手法であるクラスター分析や主成分分析においては高頻度語彙上位50語を指標として用いた分析とは異なり，Alice Bradley Sheldon作品群とErnest Hemingway作品群の分類は成功しなかった。一方，教師ありの分類手法である判別分析とSVMにおいてそれぞれ93.66%と96.49%でという正確率だった。これは，サンプルサイズを考慮に入れた分類正確率の基準を有意に上回っている。つまり，教師ありの分類手法でのみこの2著者間の分類が成功した。

2種類の指標を使用し，4種の統計手法を用いた検証により，Alice Bradley Sheldon作品群内でのJames Tiptree, Jr.名義作品群とRaccoona Sheldon名義作品群という基準による著者内変異は存在しない可能性が大きいと結論付けられる。一方で，Alice Bradley Sheldon作品群とErnest Hemingway作品群における著者間の変異は存在するということが判明した。さらに，品詞分布を指標として採用すると，この2著者間の変異は教師なしの分類手法ではとらえることができないと結論付けられる。

2.2.2　木村（2017a）

前項で述べたKimura（2016）では，Alice Bradley Sheldonの著者内変異の検証としてJames Tiptree, Jr.名義作品群とRaccoona Sheldon作品群を独立したカテゴリーとして扱った。また，

Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群を比較する際にも James Tiptree, Jr. 名義作品群と Raccoona Sheldon 名義作品群とを別のカテゴリーとして扱っていた。しかし，Alice Bradley Sheldon 作品群72作品中，James Tiptree, Jr. 名義作品群は67作品存在し，Raccoona Sheldon 名義作品群は 5 作品しか存在していない。James Tiptree, Jr. 名義作品群と Raccoona Sheldon 作品群を独立したそれぞれのカテゴリーとして分類を行うと，Raccoona Sheldon 作品群の少なさから，教師ありの分類手法を用いて分析する際にサンプルサイズの大きいカテゴリーに基づいて過学習（overfitting）が起こる可能性が存在した。そのため，木村（2017a）では，James Tiptree, Jr. 名義作品群と Raccoona Sheldon 名義作品群とを独立したカテゴリーとはせず，Alice Bradley Sheldon 全72作品と Ernest Hemingway 短編全69作品との比較を行った。

木村（2017a）では，Hirst and Feiguina（2007）や Hou and Jiang（2016）で採用されておりその有効性が示されている品詞の unigram，品詞の bigram，品詞の trigram という指標を採用した。Breiman（2001）で提唱されて金・村上（2007）において様々なジャンルのテキスト分類において，既存の分類法に比べて分類感度が高いと結論付けられているランダムフォレストを用いた。品詞分布を用いて行ったテキスト分類の結果の一部を表 1 に示す[i]。表 1 から，今回分析に使用したデータセットに対する分類正確率は93.62％だった。これは，サンプルサイズを基準とした分類正確率を有意に上回っている。また，ランダムフォレストではその近接性の計算を利用し，多次元尺度（Multi-Dimensional Scaling: MDS）法を用いて散布図を図示することができる。多次元尺度法を用いて散布図を図示した結果を図 2 に示す。

図 2 では，Alice Brandley Sheldon 作品群と Ernest Hemingway 作品群の分布を検証する。描画スペースの関係上，各作品に対して ID を振り分けた。ID1～69は Ernest Hemingway 作品群を，ID70～141は Alice Bradley Sheldon 作品群を示している。ここでは，他クラスに誤分類された作品群の ID を表示している。Ernest Hemingway 作品群においては，ID19（*Cross-Country Snow*），ID24（*Great News from the Mainland*），ID46（*The Capital of the World*），ID60（*The Short Happy Life of Francis Macomber*）という作品が誤分類されているということが確認できる。また，Alice Bradley Sheldon 作品群においては，ID70（*A Day Like any Other*），ID88（*Excusion Fare*），ID98（*I'll be Waiting for You When the Swimming Pool is Empty*），ID109（*Out of the Everywhere*），ID121（*The Man Doors Said Hello to*），ID133（*Time-Sharing Angel*）という作品が他クラスへ誤分類されているということが判明した。図 3 には Gini 係数に基づく変数の特徴度を提示した。

表 1　ランダムフォレスト出力

	Alice Sheldon	Ernest Hemingway
Alice Sheldon	**67**	5
Ernest Hemingway	4	**65**

[i] 本稿での分析には R 3.2.5 を用いた。ランダムフォレストの実行は｛randomForest｝4.6–12を使用した。

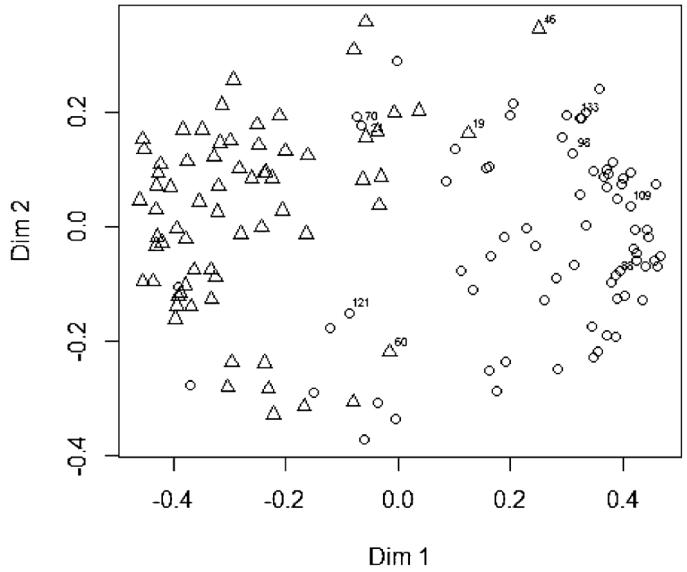

図 2　MDS 法に基づく二次元散布図

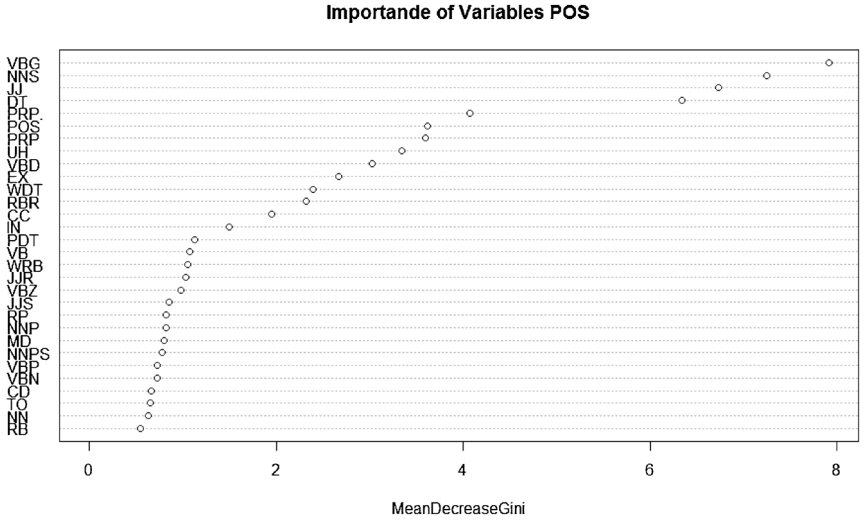

図 3　Gini 係数に基づく変数の特徴度

図 3 から，VBG（Verb, gerund/present participle），NNS（Noun, plural），JJ（Adjective），DT（Determiner）などが特徴的な変数として挙げられる。

さらに，下川・杉本・後藤（2013: 163–164）を参考にして関数を作成し，近接性に基づくクラスの中心から外れた標本の抽出である外れ値検出を行う。外れ値プロットを図 4 に示す。MDS プロット同様，外れ値プロットにおいても描画スペースの都合上 Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群をそれぞれ ID で提示する。ID1～ID69が第 1 群である Ernest Hemingway 作品群を示し，ID70～ID141が第 2 群である Alice Bradley Sheldon 作品群を示している。

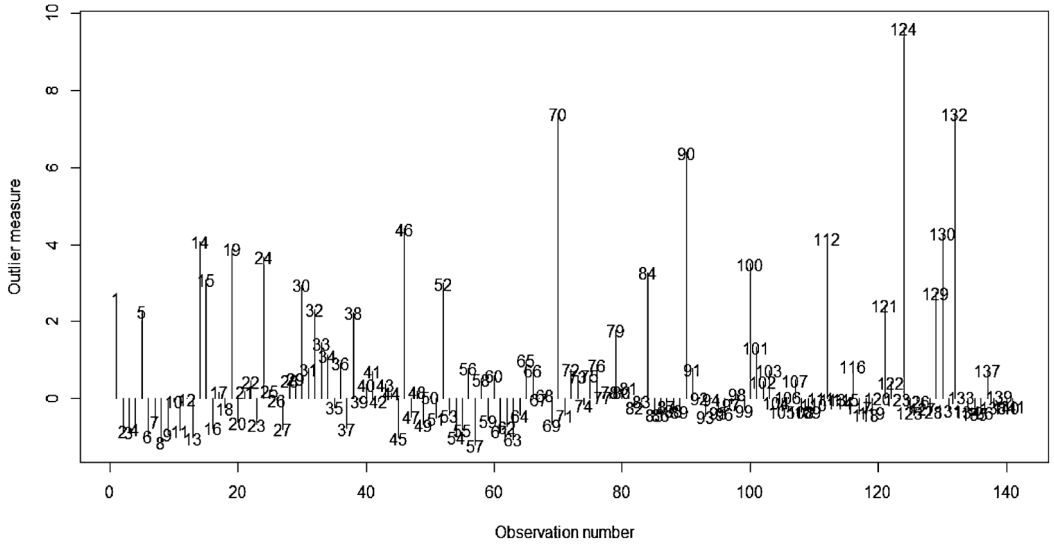

図 4　外れ値プロット

ここから，クラスの中心から外れているようなテキストの同定が可能になった。特に ID124（*The Night-Blooming Saurian*）や ID132（*Through a Lass Darkly*），ID70（*A Day Like any Other*），ID90 （*Fault*）に代表されるような外れ値の大きい作品に関して検証を行っていく必要性がある。このような検証を品詞の bigram と trigram を指標として行った結果，分類正確率はそれぞれ94.33％と90.07％であった。本研究での試行から，小谷（1994）で指摘されている Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群の主観的な文体の類似性というのは，計量文体論の手法に基づいた品詞分布の検証から来ているわけではないかもしれないと結論付けられる。

3.　データと統計手法

3.1　データ

本研究では，論文執筆者自身が構築した Alice Bradley Sheldon コーパス（延べ865,802語）を収録したコーパスから James Tiptree, Jr. 名義作品群のみを使用し，Ernest Hemingway 作品群との文体比較を行う。分析に使用した指標は，品詞の unigram，品詞の bigram，品詞の trigram である。品詞情報の付与は GoTagger[ii] というソフトウェアを使用して行った。表 2 に GoTagger の品詞タグ例を示す。

本研究では，表 3 の12冊の紙媒体の書籍を電子化し構築した Alice Bradley Sheldon（延べ865,802語）から，Raccoona Sheldon 名義作品群 5 作品を除外したデータを使用した。Alice Bradley Sheldon コーパス構築に使用した底本を表 3 に示す。

[ii] GoTagger とは，Eric Brill 氏が開発した Brill Tagger による品詞タグ付けを，Windows 上で作業可能にした GUI アプリケーションである。

表 2　GoTagger 品詞タグ一覧

品詞タグ略号	品　　詞	品詞タグ略号	品　　詞
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition/subord. conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund/present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd ps. sing.
NNP	Proper noun, singular	VBZ	Verb, 3rd ps. sing. Present
NNPS	Proper noun, plural	WDT	*wh*-determiner
PDT	Predeterminer	WP	*wh*-pronoun
POS	Possessive ending	WP$	Possessive *wh*-pronoun
PRP	Personal pronoun	WRB	*wh*-adverb

表 3　Alice Bradley Sheldon 底本

	作　品　名	出版年
1	*Brightness Falls from the Air*	1993
2	*Byte Beautiful: Eight Science Fiction Stories*	1985
3	*Crown of Stars*	1988
4	*Her Smoke Rose Up Forever*	1990
5	*Meet Me at Infinity*	2002
6	*Out of the Everywhere and Other Extraordinary Visions*	1981
7	*Star Songs of an Old Primate*	1978
8	*Tales of the Quintana Roo*	1986
9	*Ten Thousand Light-Years from Home*	1973
10	*The Starry Rift*	1986
11	*Up the Walls of the World*	1978
12	*Warm Worlds and Otherwise*	1975

また，Ernest Hemingway 作品群に関しては，*The Complete Short Stories of Ernest Hemingway* を底本として Ernest Hemingway 短編作品全69作品を含むコーパス（延べ271,475語）を論文執筆者自身が紙媒体から構築し，分析対象とした。

3.2 統計手法

3.2.1 サポートベクターマシン

サポートベクターマシン（SVM）とは，Diederich et al（2003）や Hirst and Feiguina（2007）に代表されるように，近年の著者推定の論文で使用されることの多い比較的新しい分類器である。Bayaan（2008）では，ほかの分類器に比べて分類精度が良い場合が多いとされている。金（2007: 259–260）によると，この分類器は学習データ集合 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) を用いて，以下の式(a)によって定義される。

$$y = \sum_{i=1}^{p} w_i x_i + b = WX + b \qquad \cdots(a)$$

また，本稿で使用した R パッケージ｛e1071｝に存在しているカーネル法を用いた SVM は，カーネル関数を用いて式(b)によって定義される。

$$y = f(x) = \sum_{i=1}^{n} \beta_i K(x_i, x) + b \qquad \cdots(b)$$

SVM は，マージンを最大化するような係数を求め，式(c)のように判別を行い，超平面においてグループ間の差異を最大化するような分離を行う。つまり，マージンを最大化するということは，式(c)で示すような直線 $WX_i + b = 1$ と直線 $WX_j + b = -1$ との間隔を最大化する操作である。このような操作を行うことによって，グループ間の差異を最大化し非常に高い精度の分類が可能になっている。

$$y = \begin{cases} 1 \ \ if \ \ WX + b \geq 1 \\ -1 \ \ if \ \ WX + b \leq 1 \end{cases} \qquad \cdots(c)$$

3.2.2 ランダムフォレスト

本稿では SVM に加えて，Breiman（2001）で提唱され金・村上（2007）で文芸テキストの分類の際に SVM や *k*–近傍法などの既存の分類手法に比べて分類感度が高いと結論付けられている RF という統計手法を用いた。Tabata（2012）では，この手法を用いて文芸作品の分類の際に有効であった説明変数の提示を行っており，本研究においても Tabata（2012）同様，テキスト分類を行った後，Gini 係数の平均減少率に基づいて分類の際に有効であった変数の提示を行う。RF は集団学習の手法の一種で，金（2007：271）によると以下のようなアルゴリズムに基づく。また，そのアルゴリズムは平井（2012：193）に基づいて図 5 のように平易に図解できる。

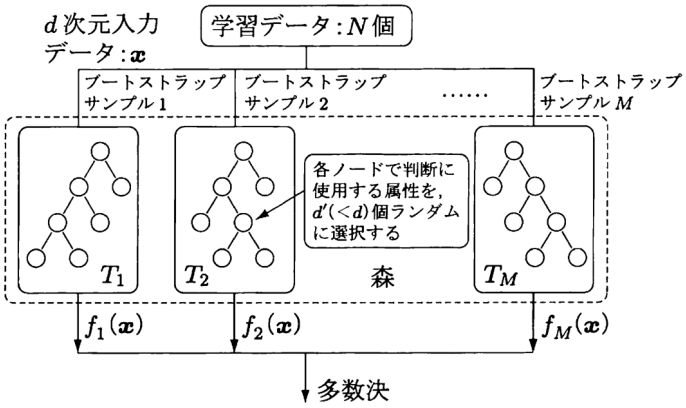

図5　決定木を用いたランダムフォレストの構成法（平井（2012：193）図11.8）

⑴　与えられたデータセットからN組のブートストラップサンプルを作成する。

⑵　各々のブートストラップサンプルデータを用いて未剪定の最大の決定・回帰木を生成する。ただし，分岐のノードはランダムサンプリングされた変数の中の最善のものを用いる。

⑶　すべての結果を統合・組み合わせ（回帰の問題では平均，分類の問題では多数決），新しい予測・分類器を構築する。

4.　分析と結果

本研究では，分類感度が高いと結論付けられているSVMとRFという2種類の統計手法と品詞unigram，品詞bigram，品詞trigramという3種の指標を用いて文書の分類を行う。分類を行った後，実行の容易さから特にRFを用いて，分類に寄与している変数の提示やその変数がJames Tiptree, Jr. 作品群とErnest Hemingway作品群のどちらに頻出しているのかということを提示する。James Tiptree, Jr. 名義作品群67作品とErnest Hemingway作品群69作品というデータセットを使用する際に6種の試行の中で，どの組み合わせが最も分類に有効であるのかという評価を行っていく。

4.1　品詞unigram

本研究では，表2で挙げた36種の変数から，頻度が0であった6種の変数を削除して分析を行った。Baayen（2008）において，他の分類手法に比べて分類感度が高いとされるSVMを用いて分析を行った。SVMの分類結果を表4に示す。

表4から，James Tiptree, Jr. 名義作品群67作品中66作品が正しく分類されているということが分かる。このクラスでの分類正確率は，98.51％だった。また，Ernest Hemingway作品群69作品はその全てが正しく分類されているということが分かる。2カテゴリーでの分類正確率は99.26％

表 4　SVM 結果

	James Tiptree	Ernest Hemingway
James Tiptree	**66**	1
Ernest Hemingway	0	**69**

表 5　RF 結果

	James Tiptree	Ernest Hemingway
James Tiptree	**60**	7
Ernest Hemingway	4	**65**

だった。これは，サンプルサイズを考慮に入れた分類正確率を有意に上回っており，この 2 カテゴリーでの分類が成功していると結論付けられる。

次に，金・村上（2007）において様々なジャンルのテキスト分類において，既存の分類法に比べて分類感度が高いとされている RF を用いた。RF の分析結果を表 5 に示す。

表 5 から，James Tiptree, Jr. 名義作品群67作品中60作品が正しく分類されて，7 作品が Ernest Hemingway 群に誤分類されているということが分かる。このクラスでの分類正確率は，89.55％だった。また，Ernest Hemingway 作品群69作品は65作品が正しく分類されて，4 作品が James Tiptree, Jr. 群に誤分類されているということが分かる。のクラスでの分類正確率は，94.20％だった。2 カテゴリーでの分類正確率は91.91％だった。この分類正確率はサンプルサイズを考慮に入れた分類正確率[iii] を有意に上回っており，この 2 カテゴリーでの分類が成功していると結論付けられる。しかしながら，RF の分類正確率91.91％は同一の変数を用いた SVM の分類正確率99.26％を下回っている。図 6 には，RF の出力から算出した MDS プロットを提示する。ここでは，誤分類されている標本に対してラベルを付けた[iv]。

図 6 では，品詞の unigram を指標として用いて分析した James　Tiptree,　Jr. 作品群と Ernest Hemingway 作品群の分布を検証する。描画スペースの関係上，各作品に対して ID を振り分けた。ここでは，他クラスに誤分類された作品群の ID を表示している。Ernest　Hemingway 作品群においては，ID32（*My Old Man*），ID42（*Summer People*），ID52（*The Good Lion*），ID60（*The Short Happy Life of Francis Macomber*）という作品が誤分類されているということが確認できる。

[iii] Kobayashi and Abe（2016）では，ランダムフォレストの分類正確率に関して“the baseline accuracy rate of the simplest possible algorithm of always choosing the most frequent category”というような基準を示している。ここでは，今回用いたテキスト全てが Ernest　Hemingway 作品群へと分類される確率を求める。その値は50.74％となる。

[iv] この機能の実装に関しては，以下の記事を参考にした。
〈http://langstat.hatenablog.com/entry/20170211/1486808584〉（参照日：2017/2/11）

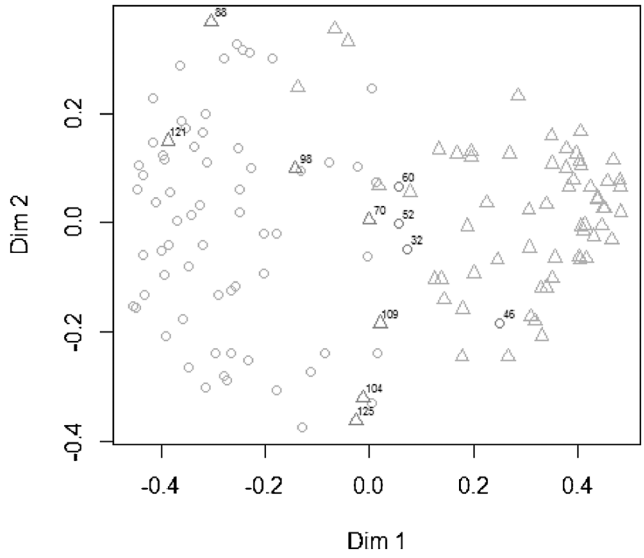

図 6　品詞 unigram MDS プロット

また，James Tiptree, Jr. 作品群においては，ID70（*A Day Like any Other*），ID88（*Fault*），ID98（*In Midst of Life*），ID104（*On the Last Afternoon*），ID109（*Press Until the Bleeding Stops*），ID121（*The Night-Blooming Saurian*），ID125（*The Snows are Melted, the Snows are Gone*）という作品が他クラスへ誤分類されているということが判明した。

判別力の高い変数とテキストとの関係性を視覚的に捉えるため，Cutler et al（2007），平井（2012），下川・杉本・後藤（2013）に基づき部分従属プロット（partial dependency plot）を図示した。部分従属プロットとは，木村（2017b）において文芸テキストの分類結果と変数との関係性について視覚化するために使用された手法であり，説明変数が目的変数に対してどう作用するのかということを提示したものである。本研究においては，説明変数である品詞分布が James Tiptree, Jr./Ernest Hemingway という 2 カテゴリーでの分類にどのように作用しているのかということを示す。描画スペースの都合上，Gini 係数の平均減少率に基づいて算出した特徴度の大きな上位10変数のみを用いる。

図 7 では縦軸の値が大きく，右肩上がりの折れ線グラフであると，その変数が第 1 群（Ernest Hemingway 作品群）において特徴的である可能性が大きいといえる。反対に縦軸の値が小さく右肩下がりになっている折れ線グラフであると，その変数が第 2 群（James Tiptree, Jr. 作品群）において特徴的である可能性が大きいということを表している。

具体的には，変数 DT は縦軸の値が小さく右肩上がりの折れ線グラフで示されているため，第 1 群である Ernest Hemingway 作品群に特徴的な変数であると結論付けられる。一方，変数 NNS は縦軸の値が大きいため第 2 群である James Tiptree, Jr. 作品群に特徴的な変数であると結論付けられる。このようにして各変数の特徴度を検証していくと，PRP, VBD という変数が第 1 群である

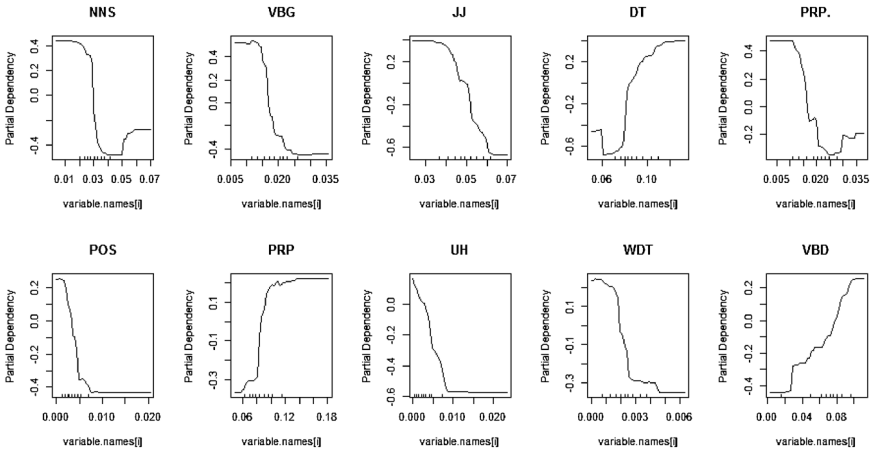

図 7　品詞 unigram 部分従属プロット

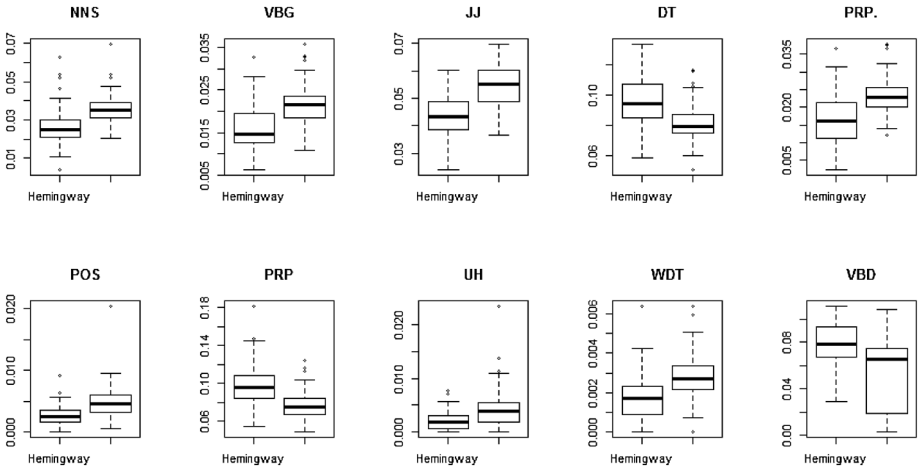

図 8　品詞 unigram 箱ひげ図

Ernest Hemingway 作品群において特徴的であり，VBG, JJ, PRP$, POS, UH, WDT が第 2 群である Ernest　Hemingway 作品群において特徴的な変数であると結論付けられる。しかしながら，このような部分従属プロットの解釈には主観的な判断が入り込む余地が存在する。箱ひげ図を用いることによって，特徴的に出現している変数に関する頻度分布を客観的に検証できるようになる。

4.2　品詞 bigram

本研究では，変数の分散を検証し「10種以上のテキストに存在している変数」という条件に合致する変数のみを抽出して分析を行った。分析に用いた変数は121種だった。SVM の出力結果を表 6 に示す。表 6 から，誤分類は 1 作品もなく SVM の分類正確率は100％であるということが分かる。

表 6　SVM 結果

	James Tiptree	Ernest Hemingway
James Tiptree	**67**	0
Ernest Hemingway	0	**69**

表 7　RF 結果

	James Tiptree	Ernest Hemingway
James Tiptree	**62**	5
Ernest Hemingway	2	**67**

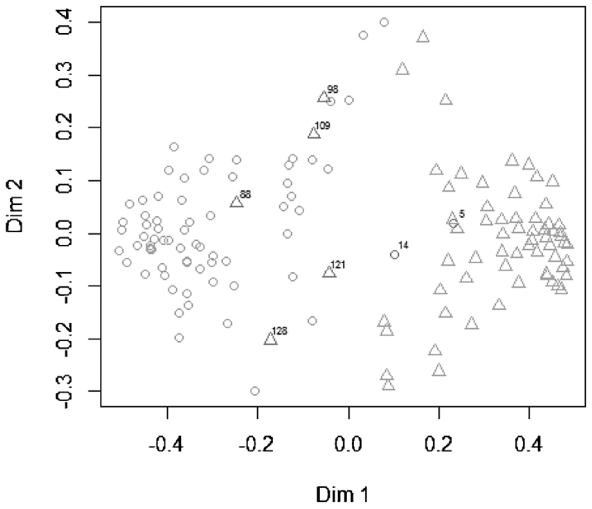

図 9　品詞 bigram MDS プロット

表 7 から，James Tiptree, Jr. 名義作品群67作品中62作品が正しく分類されて，5 作品が Ernest Hemingway 群に誤分類されているということが分かる。このクラスでの分類正確率は，92.54％だった。また，Ernest Hemingway 作品群69作品中67作品が正しく分類されて，2 作品が James Tiptree, Jr. 群に誤分類されているということが分かる。このクラスでの分類正確率は，97.10％だった。2 カテゴリーでの分類正確率は94.85％だった。これは，サンプルサイズを考慮に入れた分類正確率を有意に上回っており，この 2 カテゴリーでの分類が成功していると結論付けられる。しかしながら，RF の分類正確率94.85％は同一の変数を用いた SVM の分類正確率100％を下回っている。図 9 には，RF の出力から算出した MDS Plot を提示する。

図 9 では，品詞の bigram を指標として用いて分析した James Tiptree, Jr. 作品群と Ernest Hemingway 作品群の分布を検証する。描画スペースの関係上，他クラスに誤分類された作品群の ID を表示している。Ernest Hemingway 作品群においては，ID5（*A Natural History of the Dead*），ID14（*Banal　Story*）という作品が誤分類されているということが確認できる。また，James Tip-

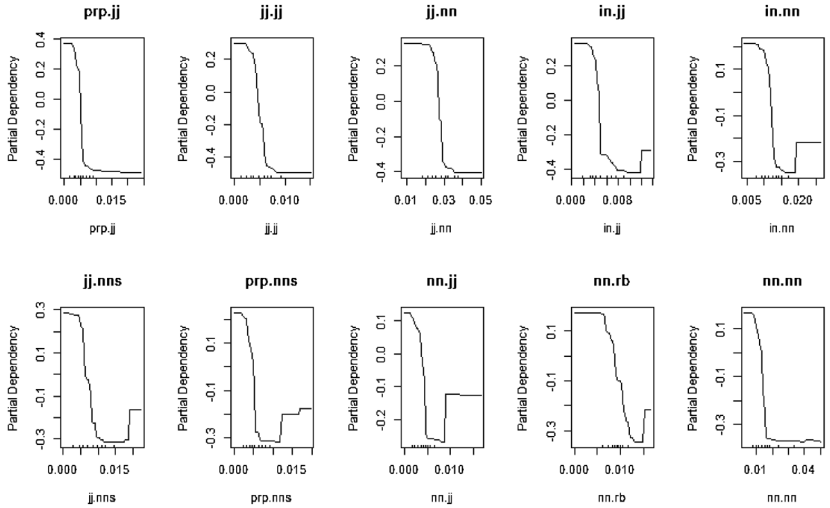

図10　品詞 bigram 部分従属プロット

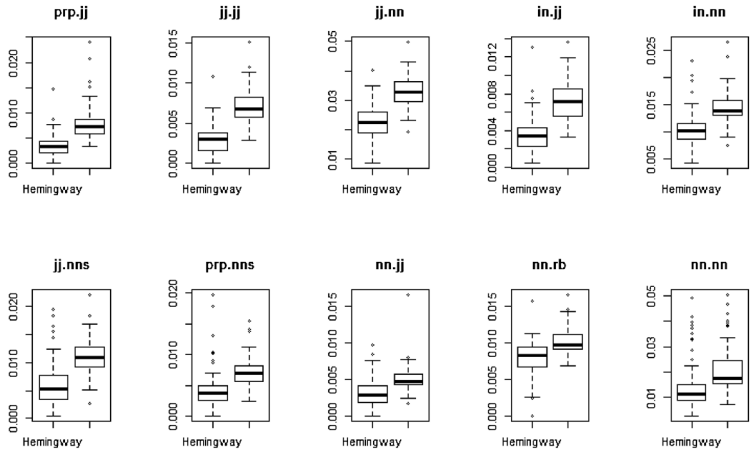

図11　品詞 bigram 箱ひげ図

tree, Jr. 作品群においては，ID88（*Fault*），ID98（*In Midst of Life*），ID109（*Press Until the Bleeding Stops*），ID121（*The Night-Blooming Saurian*），ID126（*The Trouble is not in Your Set*）という作品が他クラスへ誤分類されているということが判明した。

次に，部分従属プロットを算出した。品詞の bigram を指標として採用した分析を行う。ここでは，今回図示に用いた Gini 係数の平均減少率に基づいた特徴量の大きな変数上位10変数すべてが，第 2 群である James Tiptree, Jr. 作品群において特徴的な変数であるということが判明した。また，箱ひげ図を用いることによって，部分従属プロットで得られた結果が客観的に確認された。

4.3　品詞 trigram

本研究では，変数の分散を検証して，「10種以上のテキストに存在している変数」という条件に

合致する変数のみを抽出して分析を行った。分析に用いた変数は87種だった。

SVM の出力結果を表 8 に示す。表 8 から，James Tiptree, Jr. 名義作品群と Ernest Hemingway 作品群の分類において誤分類は 1 作品もなく SVM の分類正確率は100％であるということが分かる。

表 9 から，James Tiptree, Jr. 名義作品群67作品中60作品が正しく分類されて，7 作品が Ernest Hemingway 群に誤分類されているということが分かる。このクラスでの分類正確率は，89.55％だった。また，Ernest Hemingway 作品群69作品中63作品が正しく分類されて，6 作品が James Tiptree, Jr. 群に誤分類されているということが分かる。このクラスでの分類正確率は，91.30％だった。2 カテゴリーでの分類正確率は90.44％だった。これは，サンプルサイズを考慮に入れた分類正確率を有意に上回っており，この 2 カテゴリーでの分類が成功していると結論付けられる。しかしながら，RF の分類正確率90.44％は同一の変数を用いた SVM の分類正確率100％を下回っている。図12には，RF の出力から算出した MDS プロットを提示する。

Ernest Hemingway 作品群においては，ID5（*A Natural History of the Dead*），ID14（*Banal Story*），ID30（*Landscape with Figures*），ID37（*On the Quai at Smyrna*），ID66（*Today is Friday*），という作品が誤分類されているということが確認できる。また，James Tiptree, Jr. 作品群においては，ID70（*A Day Like any Other*），ID109（*Press Until the Bleeding Stops*），ID118（*The Man Doors Said Hello to*），ID120（*The Milk of Paradise*），ID121（*The Night-Blooming Saurian*），ID125（*The Snows are Melted, the Snows are Gone*），ID126（*The Trouble is not in Your Set*）という作品が他クラスへ誤分類されているということが判明した。

次に，部分従属プロットを算出した。変数 DT NN CC は縦軸の値が小さく右肩上がりの折れ線グラフで示されているため，第 1 群である Ernest Hemingway 作品群に特徴的な変数であると結論付けられる。一方，変数 JJ NN IN は縦軸の値が大きいため第 2 群である James Tiptree, Jr. 作品群に特徴的な変数であると結論付けられる。このようにして各変数の特徴度を検証していくと，IN IN DT という変数が第 1 群である Ernest Hemingway 作品群において特徴的であり，IN DT

表 8　SVM 結果

	James Tiptree	Ernest Hemingway
James Tiptree	**67**	0
Ernest Hemingway	0	**69**

表 9　RF 結果

	James Tiptree	Ernest Hemingway
James Tiptree	**60**	7
Ernest Hemingway	6	**63**

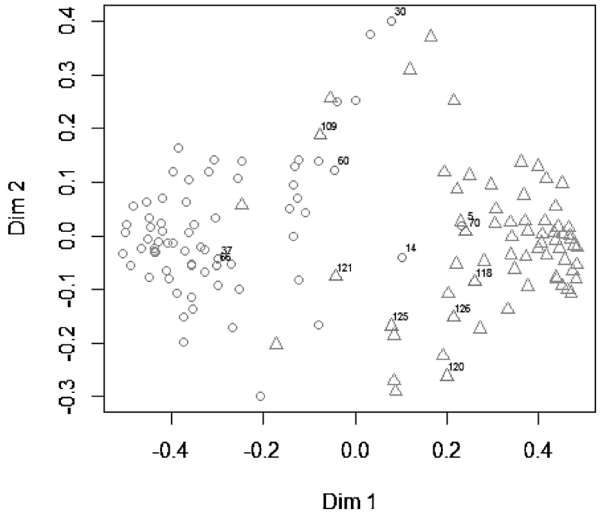

図12　品詞 trigram MDS プロット

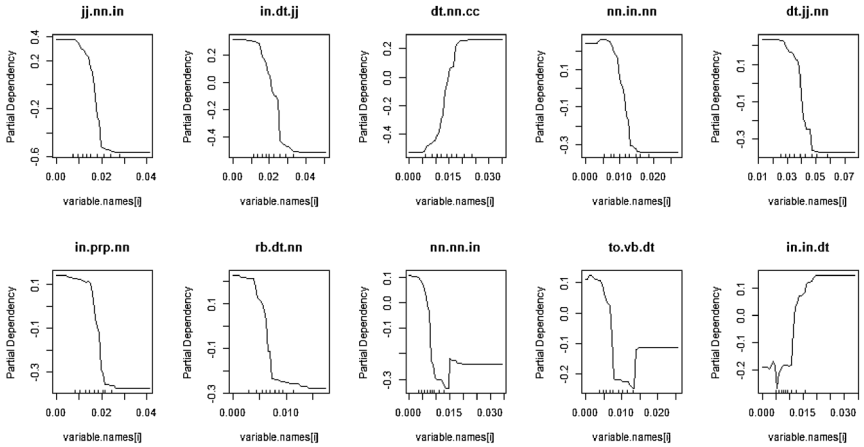

図13　品詞 trigram 部分従属プロット

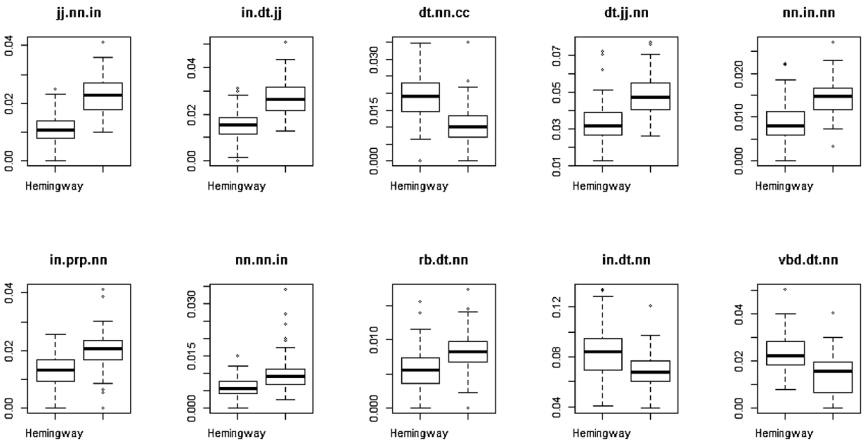

図14　品詞 trigram 箱ひげ図

JJ, NN IN NN, DT JJ NN, IN PRP NN などが第 2 群である Ernest Hemingway 作品群において特徴的な変数であると結論付けられる。また，箱ひげ図を用いることによって，特徴的に出現している変数に関する頻度分布を客観的に検証できるようになった。

5. 結　論

本研究の結論を示す。本研究では品詞分布という言語特徴を検証することによって，文芸批評上比較されることの多い作家作品群のテキスト分類を試みた。品詞の unigram を指標として分析を行った結果，James Tiptree, Jr. 名義作品群と Ernest Hemingway 作品群というデータセットを用いた場合，SVM での分類正確率は99.26％だった。同一の指標を用いて RF で分析を行った結果，この 2 群での分類正確率は92.65％だった。また，品詞の bigram を指標として分析を行った結果，この 2 群での分類正確率は，SVM は100％だった。一方，RF での分類正確率は95.59％だった。最後に，品詞の trigram を指標として分析を行った結果，SVM での分類正確率が100％である一方で，RF の分類正確率は90.44％にとどまった。Baayen（2008）で述べられているとおり，分類手法では RF よりも SVM の分類感度が高いと結論付けられる。6 種の検証結果から指標に関しては，Hirst and Feiguina（2007）で使用されている品詞の bigram が今回の試行に用いた指標の中で最も分類に有効であると結論付けられる。

小谷（1994）では主観的，印象論的な分析により James Tiptree, Jr. 作品群と Ernest Hemingway 作品群の文体の類似性が指摘されているが，SVM と RF を使用し計量文体論の手法を用いた今回の分析においてはそのような結果は追認できなかった。今後，Alice Bradley Sheldon と同時代・同ジャンルの作家作品群を含んだ大規模コーパスを構築し計量文体分析を行うことによって，この 2 著者間の文体差の検証をさらに進めていく必要性が存在する。また，部分従属プロットから，本研究で用いた 2 著者作品群は，unigram, bigram, trigram のすべての指標において形容詞原級（JJ）の使用で分類できるということが判明した。この結果から，今後の分析では James Tiptree, Jr. 作品群に特徴的に出現している形容詞に関して質的な調査を進めていく必要性がある。

謝辞

本稿の執筆に際して，何度もご指導くださった日本大学生産工学部助教小林雄一郎先生と，大阪大学大学院言語文化研究科言語文化専攻田畑智司先生に感謝の意を表します。

参考文献

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press.

Breiman, L. (2001). Random Forests, *Machine Learning*, 45, 5–32.

Burrows, J. F. (1987). *Computation into Criticism: A study of Jane Austen's novels and an experiment in method.* Oxford: Clarendon Press.

Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information, *Literary and Linguistic Computing*, 7(2), 91–109.

Burrows, J. F., & Hassal, A. J. (1988). Anna Boleyn and the authenticity of Fielding's feminine narratives, *Eighteenth Century Studies*, 21, 427–453.

Cutler, D. R., Edwards, T. C. Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. & Lawler, J. J. (2007). Random Forests for Classification in Ecology, *Ecology*, 88(11), 2783–2792.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. *Applied intelligence*, 19(1–2), 109–123.

Hirst, G., & Feiguina, O. G. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4), 405–417.

平井有三 (2012).『はじめてのパターン認識』東京：森北出版.

Hou, R., & Jiang, M. (2014). Analysis on Chinese quantitative stylistic features based on text mining. *Digital Scholarship in the Humanities*, 31(2), p. 357–367.

金明哲 (2007).『R によるデータサイエンス』東京：森北出版

金明哲・村上征勝 (2007).「ランダムフォレスト法による文章の書き手の同定」,『統計数理』, 55(2), 255–268.

Kimura, M. (2016). Can a writer disguise the true identity under pen names?: Statistical authorship attribution and the evaluation of variables, In *Proceedings of Japanese Association for Digital Humanities 2016.* 16–17.

木村美紀 (2017a).「ランダムフォレストを用いた文芸作品の分類と指標の評価—Alice Bradley Sheldon と Ernest Hemingway—」,『情報処理学会研究報告』, 2017–CH–113, 1–6.

木村美紀 (2017b).「ランダムフォレストを用いた文芸作品の計量的分類と変数の特定の試み—Alice Bradley Sheldon と Ernest Hemingway—」,『英語コーパス研究』, 第24号, 41–54.

Kobayashi, Y. & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics, 20*(1), 55–73.

小谷真理 (1994).『女性状無意識：テクノガイネーシス—女性 SF 論序説』東京：勁草書房.

Larbalestier, J. (2002). *The battle of the sexes in science fiction.* Connecticut; Wesleyan University Press.

Lefanu, S. (1988). *Feminism and Science Fiction.* Bloomington and Indianapolis; Indiana University Press.

Rybicki, J. (2015). Vive la difference: Tracing the (authorial) gender signal by multivariate analysis of word frequencies, *Digital Scholarship in the Humanities.* 31(4): 746–761.

下村敏雄・杉本知之・後藤昌司 (2013).『樹木構造近接法』東京：共立出版.

Silverberg, R. (1975). Who Is Tiptree, What Is He?, *Warm Worlds and Otherwise.* iv–xviii.

Tabata, T. (2012). Approaching Dickens' style through random forests, In University of Hamburg: *Proceedings of the Digital Humanities*: Conference Abstracts, 388–391.